DOCUMENT RESUME

ED 471 461 TM 034 627

AUTHOR Perkins, Kyle

TITLE A Scalable Set of ESL Reading Comprehension Items.

PUB DATE 2002-00-00
NOTE 35p.
PUB TYPE Reports - Research (143)
EDRS PRICE EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS *Adults; Classification; *English (Second Language); Limited English Speaking; *Reading Comprehension; *Scaling; *Test Items
IDENTIFIERS Test of English as a Foreign Language; *Unidimensionality (Tests)

ABSTRACT

Guttman implicational scaling techniques were used to identify a unidimensional set of English as a Second Language reading comprehension items. Data were analyzed from 202 students who sat for an institutional administration of the Test of English as a Foreign Language (TOEFL). The examinees who contributed to the scalable set had significantly higher TOEFL scores than those who didn't contribute to the scalable set. The distribution of native languages represented in the scalable pool was significantly different from the native language distribution of the entire sample. The scalable items had significantly fewer syllables in their question stems than the nonscalable items. The scalable item question taxonomy distribution deviated significantly from the question taxonomy distribution for all the items. The results are discussed in relation to the Linguistic Threshold Hypothesis, language transfer, capacity constrained comprehension, psycholinguistic processing approaches, universal grammar, restructuring, and the Competition Model. (Contains 1 figure, 2 tables, and 41 references.) (Author/SLD)

A Scalable Set of ESL Reading Comprehension Items

Kyle Perkins

Southern Illinois University Carbondale

Abstract

Guttman implicational scaling techniques were used to identify a unidimensional set of ESL reading comprehension items. The examinees who contributed to the scalable set had significantly higher TOEFL scores than those persons who didn't contribute to the scalable set. The distribution of native languages represented in the scalable pool was significantly different from the native language distribution of the entire sample. The scalable items had significantly fewer syllables in their question stems than the non-scalable items. The scalable item question taxonomy distribution deviated significantly for the question taxonomy distribution for all the items.

The results are discussed in relation to the Linguistic Threshold Hypothesis, language transfer, capacity constrained comprehension, psycholinguistic processing approaches, universal grammar, restructuring, and the Competition Model.

Introduction

The purposes of this study are (1) to identify a set of items in an ESL reading comprehension test that are truly scalable and unidimensional, (2) to describe the test passages using the Lexile Framework for Reading (Stenner, 1996), (3) to describe the question stems and question options using objective measurement of reading comprehension, (4) to describe the subjects in terms of their total TOEFL scores and native languages, and (5) to discuss the findings in relation to the Linguistic Threshold Hypothesis (Bernhardt and Kamil, 1995), language transfer (Grabe and Stoller, 2002), capacity constrained comprehension (Just and Carpenter, 1992), psycholinguistic processing (Snow, 1998), universal grammar (Gass and Selinker, 2001), restructuring (McLaughlin, 1990), and the Competition Model (MacWhinney, 1987; Bates and MacWhinney, 1981).

The quest to identify a set of reading comprehension items that are truly scalable and unidimensional (i.e., measure a single construct) is motivated by the wide variety of knowledge, skills, abilities, and strategies that one finds discussed in texts and articles whose general headings can be characterized as researching and teaching reading comprehension. Three quite randomly-chosen examples should be enough to make the point.

Johnston (1983) wrote that one must consider the following factors in terms of describing reading comprehension assessment tasks: " production requirements; memory and retrieval requirements; reasoning requirements; motivation; purpose; social setting and interaction; expectation and perceived task demands; and test-wiseness" (p. 34).

Omaggio (1986) listed ten factors involved in reading: "recognizing the script of a language; deducing the meaning and use of unfamiliar vocabulary; understanding information that is stated explicitly; understanding implications not explicitly stated; understanding relationships within sentences; understanding relationships between the parts of a text through cohesive devices, both grammatical and lexical; identifying the main point or the most important information; distinguishing the main idea from the supporting detail; extracting the main points in order to summarize; and understanding the communicative value and function of the text" (p. 151).

Grabe and Stoller (2002) mention lower- and higher-level processes that are engaged when we read. The lower-level processes include lexical access; syntactic parsing; semantic proposition formation; and working memory activation. The higher-level processes include text model of comprehension; situation model of reader interpretation; background knowledge and inferencing; and executive control processes (p. 27).

A review of factor analytic and multiple regression studies of reading comprehension shows an even greater variety of findings. The following list indicates the number of "factors" identified in reading measures and by whom: Davis (1944), two; Derrik (1953), three; Davis (1968, 1972), five.

Subjects

Data were analyzed from 202 students who sat for an institutional administration of a TOEFL test. The subject pool had an average score of 456.99 (SD = 59.51) on the overall TOEFL, and the distribution of native languages was as follows:

50	Chinese
35	Japanese
28	Korean
27	Spanish
13	Arabic
11	Thai
4	Cantonese
4	Malay
4	Greek
3	Hindi
3	Portuguese
3	Turkish
3	Urdu
2	Russian
2	Unknown
1	Swedish
1	Somali
1	German
1	Kurundi
1	Malinke
1	French

1	Indonesian
1	Romanian
1	Amharic
1	Mandarin

Guttman Implicational Scaling

Guttman implicational scaling was utilized to identify a scalable set of items. Guttman scaling analyzes the underlying characteristics of three or more items to determine whether the interrelationships between the items meet the properties that define a Guttman scale. Two of those properties are unidimensionality and cumulativeness.

Unidimensionality implies that items must all measure movement toward or away from the same underlying construct. Cumulativeness implies that items can be ordered by item difficulty and that subjects who "pass" a difficult item will also "pass" easy items and vice versa (Torgerson, 1958). Operationally, one looks for the extent to which scores of 1 for a given item are associated with scores of 1 for all items that have been determined to be less difficult. One also looks for the extent to which scores of 0 for a given item are associated with scores of 0 for all items that have been determined to be more difficult.

In conducting a Guttman scaling procedure, one seeks the degree to which the data fit the model. Deviations from the expected pattern are counted as errors that are aggregated, and coefficients are produced to enable the researcher to ascertain whether the items are scalable, unidimensional, and cumulative.

Four statistics are associated with Guttman implicational scaling. "The *coefficient of reproducibility* is a measure of the extent to which a respondent's scale score is a predictor of one's response pattern. It varies from 0 to 1. A general guideline to the interpretation of this measure is that a coefficient of reproducibility higher than .9 is considered to indicate a valid scale. The *minimum marginal reproducibility* constitutes the minimum coefficient of reproducibility that could have occurred for the scale given the cutting points used and the proportion of respondents passing and failing each of the items. The difference between the coefficient of reproducibility and the minimum marginal reproducibility indicates the extent to which the former is due to response patterns rather than the inherent cumulative interrelationship of the variables used. This difference is called the *percent improvement* and is actually the difference in two percents rather than a ratio itself. The final measure is obtained by dividing the percent improvement by the difference between 1 and the minimum marginal reproducibility. The denominator represents the largest value that the percent improvement may attain, and the resulting ratio is called the *coefficient of scalability.* The coefficient of scalability also varies from 0 to 1, and should be well above .6 if the scale is truly unidimensional and cumulative" (Nie, Hull, Jenkins, Steinbrenner, and Bent, 1975, pp. 532-33).

The Lexile Framework for Reading

The Lexile framework is based on the notion that all symbol systems share a semantic component and a syntactic component. "In all cases, the semantic units vary in familiarity and the syntactic structures vary in complexity. The comprehensibility or

difficulty of a message is governed largely by the familiarity of the semantic units and by the complexity of the syntactic structures used in constructing the message.

"As far as the semantic component is concerned, it is clear that most operationalizations are proxies for the probability that a person will encounter a word in context and thus infer its meaning (Bormuth, 1966). Klare (1963) builds the case for the semantic component varying along a familiarity-to-rarity continuum. Knowing the frequency of words as they are used in written and oral communication provides the best means of inferring the likelihood that a word will be encountered and thus become a part of the individual's receptive vocabulary.

"Variables such as the average number of letters or syllables per word are actually proxies for word frequency. They capitalize on the high negative correlation between the length of words and the frequency of word usage. Polysyllabic words are used less frequently than monosyllabic words, making word length a good proxy for the likelihood of an individual being exposed to them.

"Sentence length is a powerful proxy for the syntactic complexity of a passage. One important caveat is that sentence length is not the underlying causal influence (Chall, 1988). Researchers sometimes incorrectly assume that manipulation of sentence length will have a predictable effect on passage difficulty. Davidson and Kantor (1982), for example, illustrate rather clearly that sentence length can be reduced and difficulty increased and vice versa.

"Klare (1963) provides a possible interpretation for how sentence length works in predicting passage difficulty. He speculates that the syntactic component varies in the load placed on short-term memory. This explanation also is supported by Crain and Shankweiler (1988), Shankweiler and Crain (1986), and Liberman, Mann, Shankweiler, and Werfelman (1982), whose work has provided evidence that sentence length is a good proxy for the demands that structural complexity places upon verbal short-term memory" (Stenner, 1996, pp. 9-10).

MetaMetrics (MetaMetrics, 1995) computer program was used to analyze the test reading passages. The program includes sentence length and word frequency in its analysis and reports the difficulty in Lexiles. The Lexiles are anchored at the low end (200) on text from seven basal primers and at the high end (1200) on text from the *Electronic Encyclopedia* (Grolier, 1986).

Question Stem and Question Options Measures

For each question stem, the number of syllables and the average word frequency for the words appearing in the question stem were recorded. The word frequency measure used was the raw count of how often a given word appeared in a corpus of 5,088,721 words sampled from a broad range of school materials (Carroll, Davies, and Richman, 1971). The number of syllables in the question options were also recorded in addition to the average word frequency for the words appearing in the question options. Syllables are considered in this research as a proxy for syntactic complexity, and word

frequency is considered as a proxy for the likelihood that the subjects had been exposed to the word.

Question Taxonomy

A measure of the extent to which a transformation exists between a reading comprehension test passage and the options was also included in the study. Anderson (1972) produced the taxonomy of questions which was cited by Embretson and Wetzel (1987) and which was used in this research:

"1. Verbatim questions, in which a statement in the same form as the text is given as an alternative for verification;

2. Transformed verbatim questions, in which the same basic words are used in the text, but the sentences or phrases are rearranged (e.g., 'The boy hit the ball' becomes 'By whom was the ball hit?');

3. Paraphrase questions, in which the question has the same meaning as a sentence in the text, but different words are used;

4. Transformed paraphrase questions, in which neither the wording nor the phrase order in the question is the same as in the text;

5. Alternative choices that are particular instances of a superordinate term in the question stem (i.e., deduction); and

6. Questions with particular instances in the question stem and alternative choices consisting of superordinate or gist statements (i.e., induction)" (p. 176).

The Data

In summary, data from two person measures and six item measures were analyzed. The person measures included the testees' total TOEFL score and native language. The item measures included the text Lexile measures, the number of syllables in the question stem, the average word frequency in the question stem, the number of syllables in the question options, the average word frequency in the question options, and the question taxonomy.

The Analyses

The data were entered in a matrix with the subjects rank-ordered so that the examinee with the lowest number of correct responses was at the bottom of the matrix, and the examinee with the most correct responses was the top of the matrix. The items were rank-ordered in the matrix so that the item with the fewest correct responses (the most difficult item) was first, and the item with the most correct responses (the easiest item) was last. The complete matrix contained 5,858 cells (202 subjects x 29 items).

The Guttman coefficients for the entire data matrix were as follows:

Coefficient of reproducibility	.70
Maximum marginal reproducibility	.63
Percent improvement	.07
Coefficient of scalability	.19

These coefficients indicate that the original data matrix was neither scalable nor unidimensional, because the coefficient of scalability should be well above .6 if the scale

is truly unidimensional.

The following winnowing process was begun to ferret out misfitting items and misfitting persons. Phi coefficients and point biserial correlations were computed for each item. phi coefficient allows one to correlate scores for two items the scores of which are binary, i.e., correct or incorrect. The phi coefficient provides one estimate of how highly interrelated the items of a test may be. The point biserial correlation is a correlation between item responses and total scores for a test. The point biserial correlation can be used as an estimate of an item's discriminability.

Each phi coefficient was converted to a Fisher Z so that the coefficients could be averaged and then converted back to phi coefficients. The point biserial correlations were similarly converted to Fisher Z, averaged, and then converted back to point biserial correlation coefficients. Items having phi coefficients a half standard deviation below the mean were eliminated from the matrix. Items having point biserial correlations a half standard deviation below the mean were eliminated from the matrix.

Guttman person errors and Guttman item errors were identified. Figure 1 attempts to display how Guttman person and item errors were operationalized in this study. The item errors were averaged, and each item having an error score one standard deviation above the mean was eliminated. The person errors were averaged, and each person whose error score was one standard deviation above the mean was also eliminated. After a cycle of item and person eliminations had been conducted, the Guttman statistics were computed again.

The item and person elimination and Guttman coefficient calculation process was applied cyclically seven times until a coefficient of scalability of .6 was reached. The matrix was reduced from 5,858 cells (202 subjects x 29 items) to 512 cells (64 subjects x 8 items), a net reduction of 5,346 cells, to produce the following Guttman coefficients:

Coefficient of reproducibility	.93
Maximum marginal reproducibility	.82
Percent improvement	.11
Coefficient of scalability	.61

Four observations are noteworthy. First, the 202 subjects and 29 items were very multidimensional. That a reading comprehension test is multidimensional comes as no surprise to reading researchers, and that a sample of 202 students taking a TOEFL test being multidimensional comes as no surprise to second language acquisition researchers. Second, a coefficient of scalability of .61 is not well above .6, but if more subjects and more items had been eliminated to improve the scalability coefficient, there would not be a critical mass left to describe. In addition, a humane researcher wants to excise as few testees as possible. Third, a coefficient of reproducibility of .93 means that 93 percent of the time one could predict which of the eight questions a subject answered correctly based on his/her rank in the matrix. Fourth, the author has not been able to determine how far above .60 the scalability coefficient should be.

Table 1 portrays the TOEFL scores and native languages of the subjects who contributed to the scalable item set, the item numbers, the text Lexile measures, the

number of syllables in the question stem, the average word frequency in the question stem, the number of syllables in the question option, the average word frequency in the question options, and the question taxonomy.

The Mann-Whitney U-test and the chi-square goodness-of-fit test were used to test for significant differences between scalable and non-scalable person traits and item traits. A large number of tie scores were encountered while calculating the Mann-Whitney U statistics so the normal approximation with tie correction was employed. The calculation of the statistic with tie correction produces a *z* score. Table 2 presents the results.

TOEFL Scores

The average TOEFL score for the 202 subjects was 456.99; the scalable persons averaged 470.47; the non-scalable persons averaged 450.88. The scalable persons had significantly higher TOEFL scores than the non-scalable persons.

Native Languages

For all 202 subjects, the number of persons having a given native language was computed as a percentage of the entire subject pool. These percentages were then considered as a hypothesized distribution. For the 64 scalable persons, the number of persons having a given native language was computed as a percentage of 64, and these percentages were treated the scalable person distribution. The chi-square test indicated that the native language distribution of the scalable sample deviated significantly from the native language distribution of the entire sample (chi-square = 61.57, $p < .05$, df 23).

In other words, the distribution of native languages in the scalable pool was quite different from the native language distribution of the entire sample. It can be concluded that language differences made a difference in which subjects contributed to the scalable item set.

Text Lexile Measures

No significant difference obtained in the rank of Lexile measures for texts for the scalable and non-scalable items.

Syllables in Question Stems

The question stems in the scalable items averaged 16.75 syllables per stem, while the question stems in the non-scalable items averaged 23.81 syllables per stem. A significant difference obtained between the two sets of ranks. (The average number of syllables for all 29 items was 21.86). The scalable items had significantly fewer syllables in their question stems.

Word Frequency in Question Stems

No significant difference was found between the average word frequencies in question stems for the scalable and non-scalable items.

Syllables in Question Options

No significant difference was found for this variable.

Word Frequency in Question Options

No significant difference was found for this variable either.

Question Taxonomy

For each of the six question taxonomies, the number of questions manifesting a given taxonomy was computed as a percentage of the 29 questions. These percentages were then considered as a hypothesized distribution. For the eight items, the number of items manifesting a given question taxonomy was computed as a percentage of eight, and these percentages were treated as a scalable question taxonomy distribution. The chi-square test indicated that the scalable item question taxonomy distribution deviated significantly from the question taxonomy distribution for all the items (chi-square = 34.58, $p < .05$, df 5). An inspection of Table 1 indicates a predominance of taxonomy types 4 and 6.

Discussion

English language proficiency, language transfer, demands that structural complexity place on working memory, and the relationship that exists between a reading passage and a question's options (question taxonomy) appear to have exerted a significant impact upon the results. Each of these factors will now be discussed in relationship to the models and hypotheses listed in the introduction.

Linguistic Threshold and Linguistic Interdependence Hypothesis

One finding from the research reported in this paper is that examinees who contribute to the scalable reading comprehension item set had significantly higher TOEFL scores than those examinees who did not contribute to the scalable item set. It would seem that a certain threshold must be reached in second language proficiency in

order to provide a set of responses to a reading comprehension test that are scalable, unidimensional, and cumulative. The issue of a linguistic threshold has a long history in the second language research community.

The relationship between second language reading as a language problem and second language reading as a reading problem has been under investigation for nearly two decades. Alderson (1984) sought to determine whether first language reading or second language would account for the most variance in second language reading performance. Cummins (1979, 1991) distinguished between academic and cognitive language proficiency. Cummins' research addressed a continuum of language proficiency on which second language learners require more time to acquire a target language for academic purposes and less time to acquire a target language for basic communicative purposes. Clarke (1979) coined the phrase "short circuit hypothesis" to suggest that second language learners must reach a criterion or threshold level of second language proficiency before they can read a second language with facility. Cziko (1980), like Clarke, studied good and poor readers in their first language and examined their second language reading behavior.

Bernhardt and Kamil (1995) continued this genre of research by investigating what they called the Linguistic Threshold Hypothesis and the Linguistic Interdependence Hypothesis. In their formulation of the Linguistic Threshold Hypothesis, Bernhardt and Kamil stated that one must know the target language in order to read it. Cummins' research on bilingual elementary school children indicated that reading comprehension

knowledge, skills, and abilities once acquired seem to transfer across languages. These findings lead Bernhardt and Kamil to hypothesize that reading and writing are transferable and will be available in the target language, once language operations have been acquired. The Linguistic Interdependence Hypothesis was stated as follows: "Reading performance in a second language is largely shared with reading in a first language" (p.17).

Bernhardt and Kamil administered reading comprehension tests in both Spanish and in English to students in three levels of Spanish instruction. The researchers found support for both hypotheses, noting that "a general conclusion is that reading variables account for between 10 and 16 per cent in second language reading; language proficiency accounts for 30 to 38 per cent. In other words, while language proficiency accounts for a greater proportion of the variance, first language reading also makes a significant contribution" (p.25).

In summary, second language readers must command a threshold, criterion amount of second language structure (broadly defined) and vocabulary so that they can use their first language knowledge, skills, and abilities to comprehend the second language text. If second language structure and vocabulary strain the reader's information processing capacity, a subject to be discussed later, then fewer resources remain for fluent second language reading. One may argue that a canonical set of factor can not be defined for the threshold; however, the results of this study suggest that TOEFL scores in the high 400s, the effects of language transfer, demands that structural

complexity play on working memory, and the relationship that exists between a reading passage and a question's options are loci from which to start formulating a threshold.

Language Transfer

One finding in this study is that the distribution of native languages associated with the scalable item pool was significantly different from the native language distribution of the entire sample. As Grabe and Stoller note, language transfer is not uniformly automatic. If it were, a significant difference between the two native language distributions might not have obtained.

Grabe and Stoller present a variety of reasons to explain the lack of automatic language transfer. Some native language groups pay more attention to the ends of words than other language groups. Some language groups utilize visual processing differentiately (Hanley, Tzeng, and Huang, 1999; Koda, 1997). Orthographies differ in opacity/transparency in relation to grapheme-phoneme correspondences. Grabe and Stoller mention the general topics of linguistic and processing differences, individual and experiential differences, and socio-cultural and institutional differences as factors to be considered in explaining why language transfer is not uniform. These factors should also provide some of the explanation necessary to account for language distributional differences identified in this study; however, the data to pursue these issues were not available.

Capacity Constrained Comprehension

Another finding from this study is that the scalable items had significantly fewer

syllables in their question stems than the non-scalable items. In a previous section it was noted that sentence length is a proxy for the demands that structural complexity places upon verbal working memory. Just and Carpenter (1992) proposed that cognitive capacity constrains comprehension, which explains, in large measure, the finding referenced above.

Capacity constrains comprehension as follows. There is a finite amount of activation available to support both processing and storage. Each word, phrase, clause, and meaning unit has an activation level associated with it. If an element's activation level is above the established threshold, it becomes part of working memory. If an element's activation level is below the established threshold for comprehension or integration into working memory, a portion of the activation supporting "old" elements in working memory will be internally reallocated, and these elements from which activation has been removed will be forgotten. Comprehension and storage capabilities are attenuated when task demands exceed available resources.

Several implications of capacity constrained comprehension follow. Readers having adequate working memory capacity can simultaneously store meaning units from prior sentences while processing incoming propositions. Reading comprehension tasks requiring a voluminous amount of inputs are more likely to tax comprehension and storage than those tasks requiring fewer, simpler inputs. If the resource demands of a reading comprehension task exceed the available resources, the task will fail and/or activation maintaining old elements will be deallocated, leading to displacement or

forgetting. Increases in overall second language proficiency (and possibly instruction and practice) are likely to lead to greater efficiency in reading comprehension. In this study, students having attained TOEFL scores of 470 and above may have adequate English storage capacity to accommodate question stems of 17 syllables or so reliably and predictably but are unable to deal with question stems of 24 syllables or more reliably and predictably. One characteristic of a Guttman scale is the extent to which a respondent's scale score is a predictor of one's response pattern.

Psycholinguistic Processing Approaches

The final significant finding from this study to be discussed is the fact that the scalable item question taxonomy deviated significantly from the question taxonomy distribution for all items, with a preponderance of taxonomy types 4 and 6 in the scalable set. To review, type 4 is transformed paraphrase questions, in which neither the wording nor the phrase order in the question is the same as in the text, and type 6 is questions with particular instances in the question stem and alternative choices consisting of superordinate or gist statements.

Types 1, 2, and 3, verbatim, transformed, verbatim, and paraphrase questions, all involve, to varying degrees, the recognition of the equivalence in meaning amongst two or more linguistic units. Types 4, 5, and 6, transformed paraphrase, alternative choices that are particular instances of a superordinate term in the question stem, and questions with particular instances in the question stem and alternative choices consisting of superordinate or gist statements involve, at minimum, processing of extended discourse,

bridging between new and old information, establishing relations within a proposition and/or between propositions, and forward and backward inferencing.

With the exception of question 49, a type 1 question, all other types 1, 2, and 3 questions showed varying evidence of negative skew, albeit skew is normally used to refer to tests. Skew refers to a deviation from the normal distribution in that the patterns of scores are not symmetrical. Question 49 has the lowest average word frequency in its question options, 84. There are more processing costs associated with the scalable items in terms of extended discourse, bridging, establishing relations, and inferencing. And, apparently, the subjects had attained the threshold proficiency to follow on.

Under psychological processing approaches, processing speed and ease are functions of the amount of information to be processed. These approaches acknowledge no threshold of proficiency because acquisition is a continuous response to new information. Snow (1998) claims that psycholinguistic processing approaches focus on input and processing efficiency and factors such as those described above, which affect processing efficiency.

Future Directions

Universal Grammar

The identification of any scalable set of language testing items could be researched fruitfully in relation to various major second language acquisition research themes. One such theme is the question of whether language universals are the major organizing factor in second language learning. In the context of this research, the

question becomes, are language universals a major factor in determining which items and which persons contribute to a scalable item set.

Gass and Selinker (2001) claim that universals can affect a learner's interlanguage grammar in three ways: "(a) They could absolutely affect the shape of a learner's grammar at any point in time. (b) They could affect acquisition order whereby more marked forms would be the last to be acquired, or, in the case of implicational universals, one could expect fewer errors in the less marked forms. (c) They could be one of many interacting forces in the determining the shape of the learners' grammars " (p. 43).

There are diametrically opposing positions in the second language research community as to whether an adult second language learner has access to universal grammar. Bley-Vroman's (1989) Fundamental Difference Hypothesis holds that child first language acquisition and adult second language acquisition are different: children have access to universal grammar; adults don't. White (2000), on the other hand, believes that adult second language learners do have access to UG. According to White, access to UG comes in five flavors: full transfer/partial (or no) access; no transfer/full access; full transfer/full access; partial transfer/full access; and partial transfer/partial access.

Yet another intriguing candidate for further research with scalable item sets is the Subset Principle that predicts that the learner's first choice is to assume a smaller grammar that is a subset of a larger grammar. This idea is certainly appealing in a context in which 21 items were excluded from a sample of 29 items in order to have a

unidimensional, scalable, cumulative set.

Restructuring

Determining how restructuring influences the inclusion of both persons and test items into a scalable set is a promising, future research project. As Young and Perkins (1995) noted, restructuring is the reorganization of existing knowledge, or the transformation of one kind of knowledge structure into another. During restructuring, each newly-acquired datum is interpreted vis-à-vis how the learner's knowledge structures are organized at that point in time.

According to McLaughlin (1990), the sub-skills that are necessary to complete a task become automatic as the result of practice. Routinization becomes possible through the initial use of controlled processes that require conscious attention and that consume time. Controlled processes are freed up to being reallocated to other levels of processing, as the sub-skills become routinized. McLaughlin also noted that restructuring can also be co-terminous with discontinuous or qualitative changes, with each novel change constituting a new internal organization of knowledge structures and not merely the accretion of new structural elements. Development, then, can be thought of as occurring in stages, and changes can be discontinuous (Perkins, Brutten, and Gass, 1996). It follows then that modularity could be engaged to describe restructuring. One could assume, for example, that it is entirely possible that development in one domain such as orthographic processing could advance differentially from development in another domain such as determining the main idea of a text.

A recurring theme in restructuring accounts is reference to resources, capabilities, and the reallocation of processing resources that lead to restructuring. McLaughlin stated that "from an information-processing perspective, the mechanisms of change involved in restructuring result from the [learner's] developing capacities " (p.120). Mislevy (1993) defined a learner's knowledge as a "complex constellation of facts and concepts, and the networks that interconnect them, of automatized procedures and conscious heuristics...; of perspectives and strategies, and the management capabilities by which the learner focuses his efforts" (p. 28).

The Competition Model

The Competition Model (Bates and MacWhinnney, 1981; MacWhinney, 1987) seems to be an ideal explanatory model to account for how persons and items come to be included in a scalable, unidimensional, cumulative set. In the Competition Model view, language processing is seen as a series of competitions between lexical items, phonological forms, and syntactic patterns. To learn language forms, a learner must have multiple accurate exposures to patterns and words in different contexts. If given patterns and words are consistently available in the input, they are acquired. If, on the other hand, other patterns and words occur rarely, or are accompanied by white noise, they are learned late, if at all.

Second language learners face conflicts between first language and second language cues and cue strengths. To resolve such conflicts, learners first avail themselves to meaning-based cues. Finally, they gradually adopt the appropriate second language

biases as their attained second language proficiency increases. Relevant cues must be identified and readjusted and measures of the relative strengths of these cues taken. The last five sentences constitute a gross oversimplification of how the Competition Model works, but its implications for applications to scalable item studies are quite transparent.

A Final Word

Identifying a scalable set of reading comprehension test items may seem like arcane, leisure activity for language testing researchers, but it isn't. Some models require unidimensional data, and some factor analytic techniques used to study dimensionality are fraught with danger and with possible misinterpretation. Guttman implicational scaling can isolate unidimensional data, and those data can be beneficial for second language acquisition and second language reading comprehension research. Hopefully, this paper has demonstrated both.

References

Alderson, J. C. (1984). Reading in a foreign language: A reading problem or a language problem? In J. C. Alderson & A. H. Urquhart (Eds.), *Reading in a foreign language.* London: Longman.

Anderson, R. C. (1972). How to construct achievement tests to assess comprehension. *Review of Educational Research, 42,* 145-170.

Bates, E., & MacWhinney, B. (1981). Second language acquisition from a functionalist Perspective: Pragmatic, semantic, and perceptual strategies. In H. Winitz (Ed.), *Annals of the New York Academy of Science Conference on Native and Foreign Language Acquisition* (pp. 190-214). New York: New York Academy of Science.

Bernhardt, E., & Kamil, M. (1995). Interpreting relationships between L1 and L2 reading: Consolidating the linguistic threshold and the linguistic interdependence hypothesis. *Applied Linguistics, 16,* 15-34.

Bley-Vroman, R. (1989). What is the logical problem of foreign language learning? In S. Gass & J. Schachter (Eds.), *Linguistic perspectives on second language acquisition* (pp. 41-68). Cambridge: Cambridge University Press.

Bormuth, J. R. (1966). Readability: New approach. *Reading Research Quarterly 7,* 79-132.

Carroll, J. B., Davies, P., & Richman, B. (1971). *Word frequency book.* Boston: Houghton Mifflin.

Chall, J. S. (1988). The beginning years. In B. L. Zakaluk & S. J. Samuels (Eds.), *Readability: Its past, present, and future*. Newark, DE: International Reading Association.

Clarke, M. A. (1979). Reading in Spanish and English: Evidence from adult ESL students. *Language Learning, 29,* 121-150.

Crain, S., & Shankweiler, D. (1988). Syntactic complexity and reading acquisition. In A. Davidson & G. M. Green (Eds.), *Linguistic complexity and text comprehension: Readability issues reconsidered.* Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Cummins, J. (1979). Linguistics interdependence and the educational development of bilingual children. *Review of Educational Research, 49,* 222-251.

Cummins, J. (1991). Conversational and academic language proficiency in bilingual contexts. *AILA Review, 8,* 75-89.

Cziko, G. A. (1980). Language competence and reading strategies: A comparison of first-language and second-language oral reading errors. *Language Learning, 30,* 101-116.

Davidson, A., & Kantor, R. N. (1982). On the failure of readability formulas to define readable text: A case study of adaptations. *Reading Research Quarterly, 17,* 187-209.

Davis, F. B. (1944). Fundamental factors of comprehension of reading. *Pscyhometrika*, *9*, 185-197.

Davis, F. F. (1968). Research in comprehension in reading. *Reading Research Quarterly* *3*, 499-545.

Davis, F. B. (1972). Psychometric research on comprehension in reading. *Reading Research Quarterly*, *7*, 628-678.

Derrik, C. (1953). Three aspects of reading comprehension as measured by tests of different lengths. *Research Bulletin 53-8*. Princeton, NJ: Educational Testing Service.

Electronic Encyclopedia. (1986). Danbury, CT: Grolier.

Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph Comprehension test. *Applied Psychological Measurement, 11,* 175-193.

Gass, S. M., & Selinker, L. (2001). *Second language acquisition: An introductory course.* Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Grabe, W., & Stoller, F. (2002). *Teaching and researching reading.* London: Longman.

Hanley, J. R., Tzeng, O., & Huang, H. S. (1999). In M. Harris & G. Hatano (Eds.), *Learning to read and write: A cross-linguistic perspective* (pp. 173-195). Cambridge: Cambridge University Press.

Johnston, P. H. (1983). *Reading comprehension assessment: A cognitive basis.* Newark, DE: International Reading Association.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99,* 122-149,

Klare, G. R. (1963). *The measurement of readability.* Ames, IA: Iowa State University Press.

Koda, K. (1997). Orthographic knowledge in L2 lexical processing. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition* (pp. 35-52). New York: Cambridge University Press.

Liberman, I. Y., Mann, V. A., Shankweiler, D., & Werfelman, M. (1982). Children's memory for recurring linguistic and non-linguistic material in relation to Reading ability. *Cortex, 18,* 367-375.

MacWhinney, B. (1987). Applying the competition model to bilingualism. *Applied Psycholinguistics, 8,* 315-328.

McLaughlin, B. (1990). Restructuring. *Applied Linguistics, 11,* 113-127.

MetaMetrics Computer Program (1995). Durham, NC: MetaMetrics, Inc.

Mislevy, R. J. (1993). Foundations of a new test theory. In N. Frederiksen, R. J. Mislevy, & I. Bejar (Eds.), *Test theory for a new generation of tests.* Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent. D. H. (1975). *Statistical package for the social sciences. Second edition.* New York: McGraw-Hill Book Company.

Omaggio. A. C. (1986). *Teaching language in context. Proficiency-oriented instruction.* Boston: Heinle & Heinle Publishers, Inc.

Perkins, K., Brutten, S. R., & Gass, S. (1996). An investigation of patterns of discontinuous learning: Implications for ESL measurement. *Language Testing, 13,* 63-82.

Shankweiler, D., & Crain, S. (1986). Language mechanisms and reading disorders: A modular approach. *Cognition, 14,* 139-168.

Snow, C. E. (1998). Bilingualism and second language acquisition. In J. B. Gleason, & N. B. Ratner (Eds.), *Psycholinguistics, Second edition* (pp. 453-481). Fort Worth: HarcourtBrace College Publishers.

Stenner, A. J. (1996). Measuring reading comprehension with the Lexile Framework. Paper presented at the Fourth North American Conference on Adolescent/ Adult Literacy. Washington, D.C., February.

Torgerson, W. S. (1958). *Theory and methods of scaling.* New York: Wiley.

White, L. (2000). Second language acquisition: From initial to final state. In J. Archibald (Ed.), *Second language acquisition and linguistic theory* pp. 130-155). Oxford: Basil Blackwell.

Young, R., & Perkins, K. (1995). Cognition and conation in second language acquisition theory. *International Review of Applied Linguistics in Language Teaching, 33,* 142-164.

Table 1

A Scalable TOEFL Reading Comprehension Item Set

StudentID	Item 59	Item 48	Item 41	Item 44	Item 35	Item 43	Item 49	Item 46	Person Score	TOEFL Score	Native Lang
11780	0	1	1	1	1	1	1	1	7	513	Chinese
11651	0	1	1	1	1	1	1	1	7	620	Swedish
49272	0	1	1	1	1	1	1	1	7	577	Russian
17808	0	1	1	1	1	1	1	1	7	550	Spanish
17913	0	1	1	1	1	1	1	1	7	527	Japanese
17946	0	1	1	1	1	1	1	1	7	467	Korean
70116	0	1	1	1	1	1	1	1	7	593	German
17666	0	1	1	1	1	1	1	1	7	457	Korean
11715	0	1	1	1	1	1	1	1	7	533	Malay
17886	0	1	1	1	1	1	1	1	7	483	Japanese
65384	0	1	1	1	1	1	1	1	7	483	Japanese
10056	0	1	1	1	1	1	1	1	7	490	Malay
17854	0	1	1	1	1	1	1	1	7	507	Japanese
17845	0	1	1	1	1	1	1	1	7	510	Thai
17804	0	1	1	1	1	1	1	1	7	473	Spanish
9123	0	1	1	1	1	1	1	1	7	497	Chinese
62561	1	1	1	0	1	1	1	1	7	550	Kurundi
17842	0	1	1	1	1	1	1	1	7	510	Arabic
17692	0	1	1	1	1	1	1	1	7	503	Japanese
11438	0	1	1	1	1	1	1	1	7	493	Chinese
17732	0	1	1	1	1	1	1	1	7	447	Chinese
17799	0	1	1	1	1	1	1	1	7	463	Japanese
17785	0	1	1	1	1	1	1	1	7	440	Chinese
9140	0	1	1	1	1	1	1	1	7	397	Korean
91269	0	1	0	1	1	1	1	1	6	543	Cantonese
11069	0	1	1	1	0	1	1	1	6	627	Hindi
9280	0	0	1	1	1	1	1	1	6	503	Japanese
10194	0	1	1	1	1	1	1	0	6	520	Chinese
17720	0	1	1	1	0	1	1	1	6	467	Korean
8129	0	0	1	1	1	1	1	1	6	577	French
832	0	0	1	1	1	1	1	1	6	493	Arabic
17994	0	0	1	1	1	1	1	1	6	427	Japanese
49273	0	1	1	1	0	1	1	1	6	573	Russian
17857	0	0	1	1	1	1	1	1	6	510	Thai
17668	0	0	1	1	1	1	1	1	6	477	Turkish
7695	0	0	1	1	1	1	1	1	6	450	Urdu
40878	0	1	1	1	0	1	1	1	6	437	Japanese
11301	0	0	1	1	1	1	1	1	6	497	Malay
17745	0	0	1	1	1	1	1	1	6	400	Arabic
10094	0	0	1	1	1	1	1	1	6	477	Chinese
17788	0	1	1	1	1	1	0	1	6	457	Japanese
17552	0	1	0	1	1	1	1	1	6	427	Chinese
17641	0	0	1	1	1	1	1	1	6	433	Korean
12645	0	0	1	1	1	1	0	1	5	563	Malinke
17494	0	0	1	1	1	1	1	0	5	443	Thai
8990	0	0	0	1	1	1	1	1	5	483	Chinese
2907	0	0	1	1	1	0	1	1	5	457	Japanese
1139	0	0	0	1	1	1	1	1	5	390	Chinese
11445	0	0	1	0	1	1	1	1	5	437	Chinese
17942	0	0	1	0	1	1	1	1	5	433	Romanian
17772	0	0	0	0	1	1	1	1	4	410	Japanese
18001	0	0	0	0	1	1	1	1	4	440	Portug.
17643	0	0	0	1	0	1	1	1	4	430	Chinese
17648	0	0	0	0	1	1	1	1	4	420	Japanese
12222	0	0	0	0	1	1	1	1	4	397	Turkish
9075	0	0	0	0	0	1	1	1	3	410	Chinese
11394	0	0	0	0	0	1	1	1	3	397	Chinese
17932	0	0	0	1	0	1	1	0	3	363	Spanish
17966	0	0	0	0	0	0	1	1	2	413	Japanese
17620	0	0	0	0	0	0	1	1	2	397	Arabic
17944	0	0	0	0	0	0	1	1	2	343	Korean
17995	0	0	0	0	0	0	1	1	2	373	Japanese
17934	0	0	0	0	0	0	1	1	2	340	Spanish
17945	0	0	0	0	0	1	0	0	1	393	Turkish
Item Score	1	32	46	49	50	59	60	60	357		
Lexile	1270	1560	1270	1140	970	1270	1560	1140			
SylQStem	26	7	20	15	18	20	16	12			
WFrqQSt	69835	72816	70672	125873	130036	94624	52990	21207			
SylQOpt	81	127	12	33	39	15	6	8			
WFrqQOpt	32160	34593	21691	71401	20177	1365	84	46			
Qtax	6	4	6	6	5	4	1	4			

Lexile=Text Lexile Measure
SylQStem=Number of syllables in question stem
WFrqQSt=Average word frequency in question stem
SylQOpt=Number of syllables in question options
WFrqQOpt=Average word frequency in question options
Qtax=Question taxonomy

Table 2

Average of Scalable and Non-Scalable Persons and Items

	Scalable	Non-scale	Test Stat.
TOEFL	470.47	450.88	z = -1.82, p < .05, df 200
SylQStem	16.75	23.81	z = -2.16, p < .05, df 27
WFrqQSt	82122.78	73617.44	
SylQOpt	40.13	32.38	
WFrqQOpt	32651.79	34892.89	

Figure 1

Guttman Person Errors

		Items						Person Scores
		3	1	4	2	5	6	
Subjects	727	0	1	1	1	1	1	5
	707	0	1	1	1	1	1	5
	717	0	0	1	1	1	1	4
	757	0	0	0	1	1	1	3
	767	0	0	0	0	1	1	2
	1011	0	0	0	0	0	1	1
Item Scores		1	2	2	4	5	6	

There are two person errors associated with examinee 727 and they are circled. Examinee 727 should have answered item 4 correctly and should have answered item 3 incorrectly, IF this were a perfect Guttman scale.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

TM034627

I. DOCUMENT IDENTIFICATION:

Title: A Scalable Set of ESL Reading Comprehension Items	
Author(s): Perkins	
Corporate Source:	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Tyler Perkins	Printed Name/Position/Title: Associate Provost
Organization/Address: Southern Illinois University Carbondale, IL 62901-4305	Telephone: 618-536-6607 FAX: 618-453-4710
	E-Mail Address: tperkins@siu.edu Date: 11/8/02

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org